LLAVIDAL : A Large LAnguage VIsion Model for Daily Activities of Living

Supplementary Material

A. Overview

The Supplementary material is organized as follows:

- Section B: Related Work
- Section C: Additional Dataset Details
- Section D: Additional Implementation Details
- Section E: Improving Actions: Skeleton Cues vs Object Cues
- Section F: Additional Qualitative Evaluation
- Section G: LLM Prompts Used
- Section H: Limitations
- Section I: Licensing and Intended Use

B. Related Work

In this section, we discuss the recent datasets proposed for instruction tuning of LLVMs. We also present the recent advancements in multimodal conversational models both with training-free methods and methods using visual connectors.

Data: Existing video-instruction datasets, such as VideoChat[35], Valley[44], Video-ChatGPT [45], and TimeIT [56], have made significant strides in advancing general video understanding and dialogue. Valley is derived from a website called Jukinmedia that provides videos with diverse categories and wide detailed descriptions. TimeIT dataset from TimeChat offers videos with temporal variations and task diversity. On the other hand, ActivityNet [8] boasts a diverse taxonomy with 203 activity classes, most activity classes are not tailored to the ADL domain. It is to be noted that most LLVMs like VideoChatgpt [45], VideoLlava [38] derive their instruction dataset from ActivityNet. Webvid, which is now de-comissioned due to privacy issues introduced in [3], consists of 2.5 million video-text pairs scraped from the web. Although these video instruction datasets are large in scale, offer diverse action classes, and focus on general video understanding, they fail to address the challenges specific to ADL. These challenges include intra-class temporal variations, long-term temporal relationships, complex human-object interactions, and videos captured in multiview settings. Unlike ADL videos, the internet videos in these datasets are predominantly shot by a cameraman, ensuring human-centered frames. Consequently, they do not capture the unstructured randomness spatially and temporally inherent in real-world ADL videos. In contrast, ADL-X is specifically designed to address the challenges inherent in ADL (see Table 5). It captures temporal unstructuredness through the temporal stitching of several unrelated actions in sequence and incorporates complex human-object interactions from the NTU120 dataset. Additionally, our proposed PAG and WS video description techniques effectively eliminate hallucinations, resulting in high-quality video-instruction pairs.

Image captioners + LLM. Advancements in Large Language Models (LLMs) have naturally extended vision-language models [52], leveraging LLMs to enhance reasoning capabilities. The emergence of these foundation models has given rise to training-free methods like Socratic Models [75] and VideoChat [35], which use pretrained vision encoders [34, 70] to map visual information into a language embedding space, followed by LLMs for downstream video tasks.

On the other hand, effective image captioners like CogVLM [68] introduce separate layers into the Transformer block of the LLM to process image features using independent QKV matrices and Feed Forward Networks specifically designed for images. Such effective captioning approaches have inspired methods that map visual information to language via image captioners, followed by processing with LLMs. Among dialog-based models, VideoChat-Captioner [10] summarizes videos through conversations between ChatGPT [7] and a captioner such as BLIP2 [33]. Similarly, ChatVideo [67] employs task-specific foundation models to create a database of ”tracklets,” with a database manager and ChatGPT [7] collaborating to generate responses for user queries during inference.

For long video understanding, approaches like [71, 77] segment videos into smaller units, providing either segment-level descriptions directly to LLMs or encoding each segment, concatenating tokens, and projecting them into the LLM space. Likewise, Language Repository [29] introduces write-and-read operations to prune text redundancies and extract information across various temporal scales. The Multimodal Video Understanding Framework [53] explores integrating video-specific information into an LLM-based framework by using off-the-shelf vision tools to extract three object-centric modalities from videos and fusing this information through natural language. Additionally, [51] investigates optimal strategies for key-frame selection to significantly reduce redundancies. However, despite these advances, training-free models fail to capture the complex temporal relationships intrinsic to ADL. These

Table 5. Video Instruction Dataset Comparison.

Dataset	Modalities	Subjects	Multiple Views	Videos	QA Pairs	Atomic Actions per Vid	Temporal Rand.	Object Traj.	Type
TimeIT[56]	RGB+L	NA	No	173000	173K	Medium	No	No	Web
VideoChat[35]	RGB+L	NA	No	8196	11K	Low	No	No	Web
Valley[44]	RGB+L	NA	No	64,687	65K	Low	No	No	Web
VideoChatGPT [45]	RGB+L	NA	No	27,801	100K	Medium	No	No	Web
ADL-X	RGB+S+L	106	Yes	16,343	100K	High	Yes	Yes	ADL

relationships, including long-term dependencies and intricate human-object interactions, remain a significant challenge for these approaches.

Large Language Vision Models (LLVMs). The abilities of LLMs in contextual understanding and language generation have led to the introduction of video conversational models. These methods typically employ foundation models to extract visual features from images and project them into an embedding space compatible with language models. Flamingo [1] utilizes a vision-language resampler combined with gated cross-attention, while BLIP2 [33] introduces Q-Former to map image features into the LLM embedding space. Similarly, MiniGPT4 [81] uses a simple linear projection layer. However, these models fall short of becoming conversational assistants due to the lack of human instruction feedback. To address this, mPLUG-OWL [74] first aligns visual and language features through multimodal autoregressive pretraining, followed by multimodal instruction tuning using LoRA [25], enabling more natural and human-aligned responses. Models such as InstructBLIP [17] and LLaVA [40] introduce large-scale human instruction datasets to facilitate LLM fine-tuning. Meanwhile, models like PaLI [12] and Qwen-VL [2] allow direct training of LLMs during pretraining or supervised fine-tuning stages.

Other models, including VideoChat, VideoLLaMA, and TimeChat [35, 56, 78], leverage Q-Former for effective feature encoding and alignment. For example, VideoLLaMA [78] employs a vision transformer with an image Q-Former to obtain frame-level representations, followed by a video Q-Former for temporal modeling. Similarly, TimeChat [56] encodes variable-length videos using a timestamp-aware frame encoder with a Q-Former to infuse temporal information into vision tokens, followed by a sliding window Q-Former to condense frame-level features for the projection layer. Building on these approaches, VideoLLaVA [38] jointly trains on both images and videos, pre-aligning visual modalities to language using LanguageBind [80] encoders. VideoChatGPT [45] leverages both temporal and spatial features from videos, obtained by average pooling frame-level features both spatially and temporally.

In contrast to these models, LLAVIDAL incorporates 3D skeleton data and human-object interaction (HOI) cues alongside videos into the LLM embedding space. This integration enables the additional cues to learn discriminative video representations, making LLAVIDAL particularly effective for interpreting ADL videos, where temporal relationships and complex HOIs are crucial.

C. Additional Dataset Details

Question Types. We divide our QA in different questions so that our model understands human object interaction holistically, we lay emphasis on actions performed and the sequence of actions occurring in the video and likewise how objects are associated with the actions. We carefully design such questions relevant to the videos with GPT 3.5 Turbo. The questions encompasses *actions happening, summarization, objects in the scene, color of the objects and questions related to the video*. For Skeleton as QA and Object as QA, we construct two additional questions for each. For Skeleton, we include *"What is the motion of the body and joints relative to the actions?"* and *"Which joints are moving in the video?"*. For Object, we add *"What are the relevant objects in the scene?"* and *"What is the object in the trajectory [x1, y1, x2, y2]?"*. These are illustrated in Figures 6 and 7.

Average video and sentence length. There is an average of 23 words per sentence in our QA and average word count for each answer is 42. The average video length is 10 seconds in our dataset. We have $126,2229$ nouns, $551,172$ verbs, $40,415$ actions and $722,807$ objects in our QA showing the overall dynamics of dataset which is illustrated in WordCloud of the Figure 8.

Importance of Person Augmented Generation (PAG). When PAG is not applied to the videos in an ADL dataset, the resulting dense-level captions often include a significant amount of irrelevant information about the background scene. This extraneous information is not directly connected to the subject or the actions being performed, and its presence can introduce noise into the training data. If left unchecked, this noise can have a detrimental effect on the learning process, as the model may erroneously focus on the background details rather than the key elements of the ADL. By failing to isolate the relevant information, the model's attention is diverted away from the crucial aspects of the task at hand, namely the individual performing the ac-

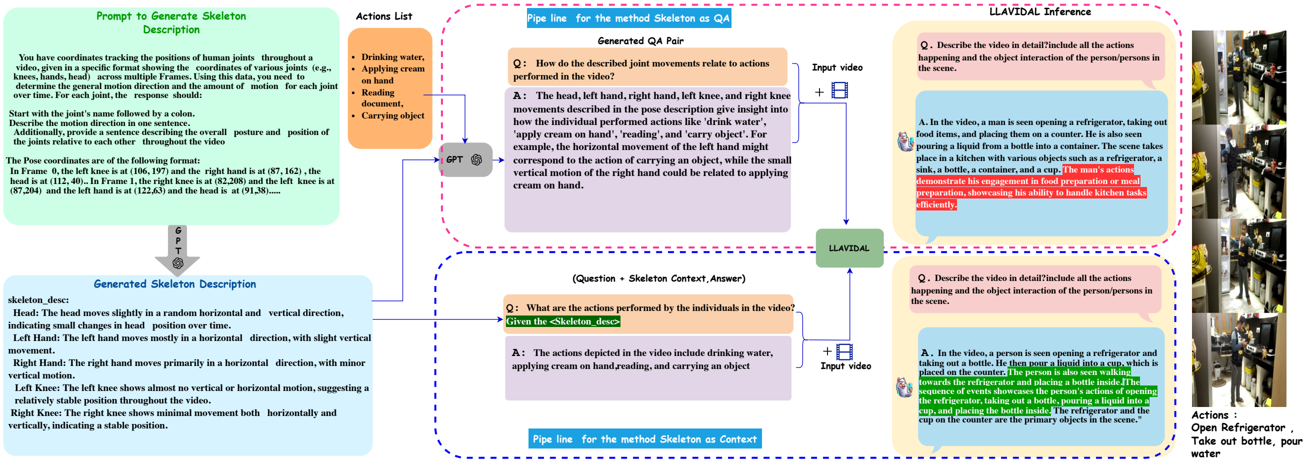

Figure 6. Overview of pipeline for Skeleton as QA and Skeleton as context.

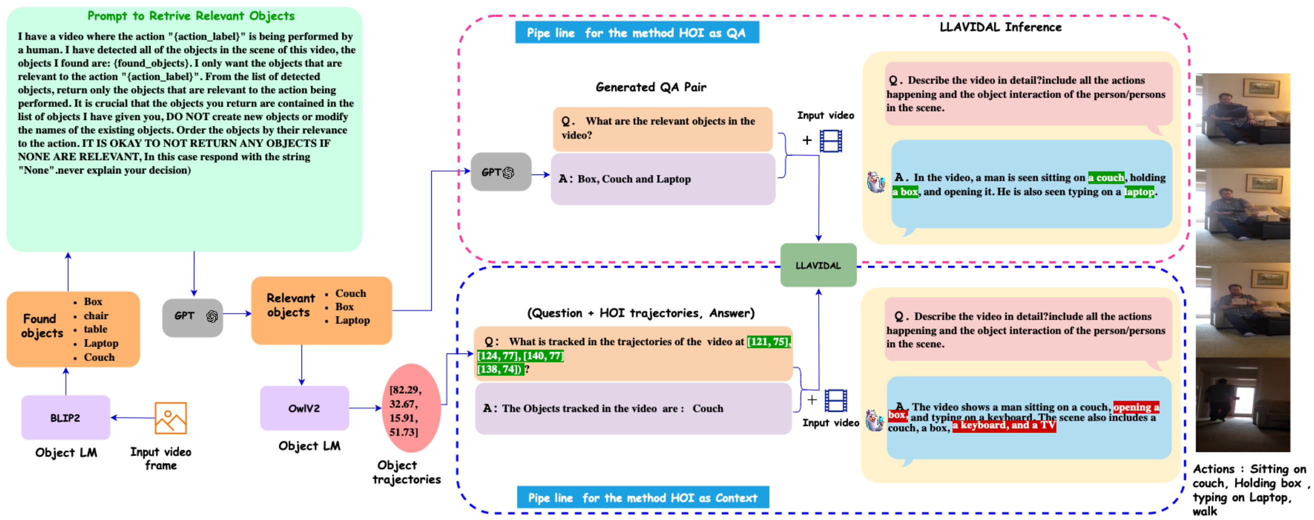

Figure 7. Overview of pipeline for HOI as QA and HOI as context.

tions, the actions themselves, and the interactions between the person and objects in the scene. This dilution of focus can lead to suboptimal performance and hinder the model's ability to accurately understand and classify ADLs. In contrast, by employing person-centric cropping, the irrelevant background information is effectively eliminated from the videos. This targeted approach ensures that the dense-level captions concentrate solely on the elements that are directly related to the subject and their actions. By maintaining this persistent focus on the relevant information, the training data becomes more coherent and informative, enabling the model to better capture the essential characteristics of the ADLs. In Fig 9, we illustrate an example to highlight the importance of PAG in our semi-automated data curation framework.

D. Additional Implementation Details

We deployed a 4-bit quantized version of CogVLM-17B [68] for annotating frame-level captions. On an A5000 GPU, the inference uses 11GB of memory.The two prompts that are used to get the frame-level descriptions for the ADL-X are – *"Give a detailed description of the actions happening and describe the image, include motions and the objects interacted by the person"* and *"Summarize the content of the image in details explaining all events happening"*. CogVLM uses Vicuna v1.5 7b [13] as their large language model and EVA2-CLIP-E [64] as their VIT encoder, the input image dimensions are 224×224, the average time to annotate a video is *80 seconds* at *0.5fps*.

LLAVIDAL details. To generate HOI cues, we perform frame-level object detection using BLIP2 and localization using OWLv2. BLIP2 [33] uses a ViT-L and a FlanT5 [15]

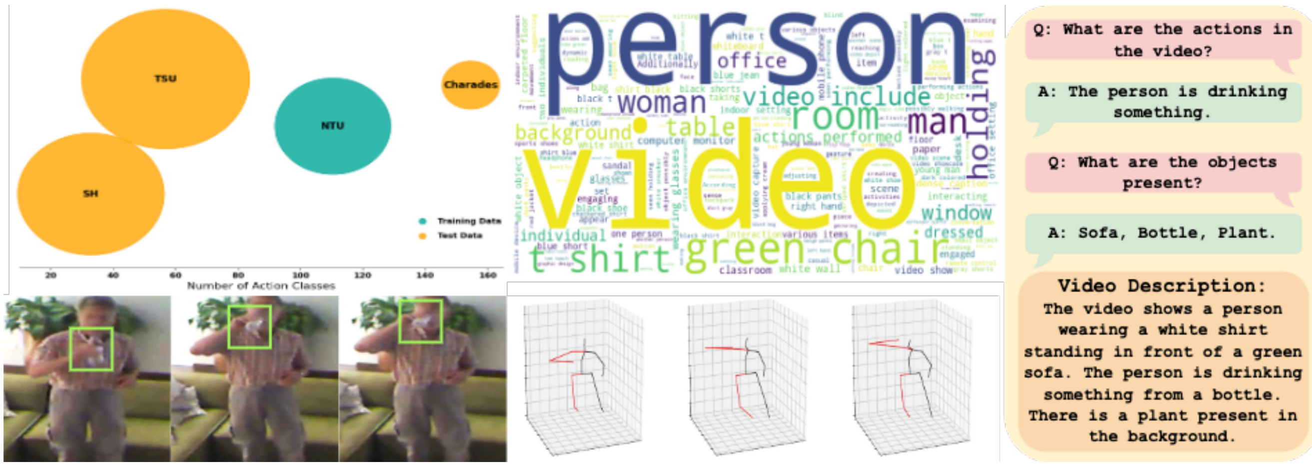

Figure 8. Overview of ADL-X. Top Left: Training and test data distribution; Top Middle: Wordcloud of Textual Representation of Training Data; Bottom Left: Sample video frames with detected relevant object Bottom Middle: 3D skeletons of the corresponding sample video; Right: Sample QA pairs

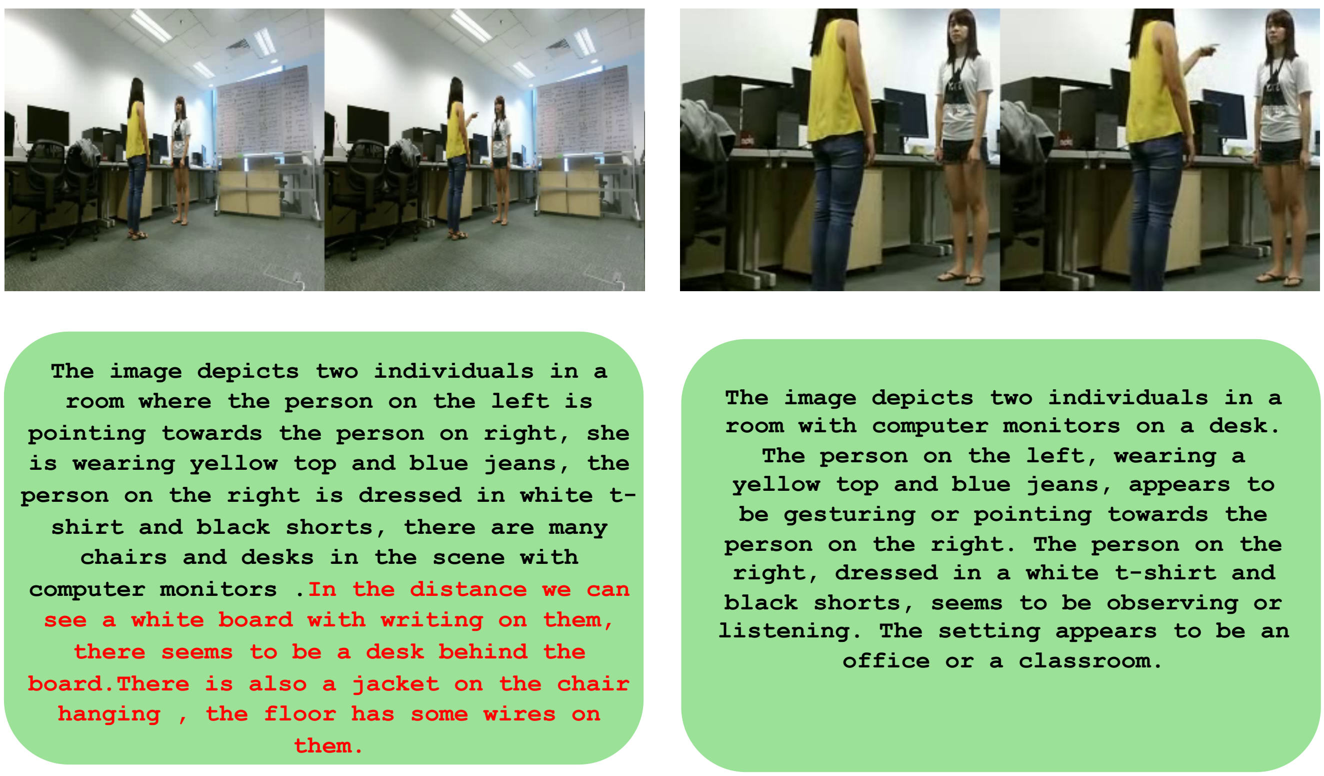

Figure 9. Left: uncropped videos and frame level annotations from CogVLM; Right: PAG and CogVLM captions. The irrelevant information (marked red) adds noise to the annotations.

architecture for detection, while OWLv2 [47] uses an OWL-ViT-L which is a CLIP based model for extracting localization features of the detected objects. In case of skeleton-CLIP, the skeleton Encoder, Hyperformer, is pretrained on NTURGBD for 140 epochs for action recognition, and then is aligned with the CLIP Text Encoder for an additional 100 epochs. LLAVIDAL uses a Vicuna-v1.1 (7B) as the LLM which is frozen during instruction tuning.

E. Improving Actions: Skeleton vs HOI Cues

In this section, we present a comprehensive analysis of our multi-modal approach leveraging HOI tokens, skeleton tokens, and their progressive integration (MMPro) for fine-grained action recognition. Our empirical evaluation across a diverse set of household actions demonstrates distinct performance patterns across modalities, revealing their complementary nature in action understanding when compared against video only trained model. This analysis is shown in Figure 10.

HOI token integration along with video token in LLMs yields substantial performance gains in object-interaction intensive actions, with peak improvements in "use laptop" (+45%), "use tablet" (+32%), "cook cut" (+40%), and "make coffee" (+35%). This performance boost can be attributed to the enhanced object-contextual reasoning capabilities, where the model effectively leverages object-centric features to disambiguate actions occurring in similar spatial contexts. The HOI token cue demonstrates particular efficacy in scenarios requiring fine-grained object state understanding, such as distinguishing between "pour from kettle" (+20%), "pour from can" (+18%), and "pour from bottle" (+15%), where object state transitions and object-specific attributes are crucial for action classification. Additionally, HOI tokens show significant improvements in context dependent actions like "breakfast eat at table" (+25%), "put something on table" (+22%), and "clean dishes dry up" (+18%), where spatial relationships between multiple objects exists.

The skeleton token cue exhibits superior performance in actions characterized by distinctive kinematic patterns, showing significant improvements in body-centric actions such as "take pills" (+28%), "drink from can" (+25%), "drink from bottle" (+22%), and "walk" (+15%). These improvements stem from the model's ability to capture fine-grained skeletal dynamics, enabling robust discrimination between actions with similar object interactions but distinct motion patterns. Notably, skeleton tokens demonstrate enhanced capability in temporal action modeling, particularly in sequential actions involving multiple body positions such as "lay down" (+18%), "get up" (+16%), and "sit down" (+17%). The skeleton cue also excels in capturing subtle motion differences in drinking actions ("drink from cup" +20%, "drink from can" +25%, "drink from bottle" +22%), where the trajectory and orientation of movement are key discriminative features.

Our proposed MMPro framework, demonstrates significant synergistic effects, particularly in complex actions requiring both object and kinematic understanding. For instance, MMPro achieves significant improvements in "use tablet" (+58%), "make coffee" (+52%), "breakfast take ham" (+45%), and "use laptop" (+48%), where both object state changes and body motion patterns are crucial for accurate classification. The framework's effectiveness is particularly evident in ambiguous scenarios where individual modalities underperform, such as "clean dishes" (HOI: +18%, skeleton: +15%, MMPro: +25%), "cook stir" (HOI: +22%, skeleton: +20%, MMPro: +32%), and "pour from kettle" (HOI: +20%, skeleton: +22%, MMPro: +30%).

In complex composite actions like "make coffee" (involving "pour grains" +28%, "pour water" +25%), MMPro successfully captures both the object state transitions and the associated body movements, resulting in more accurate action classification. Our results demonstrate that the MMPro strategy successfully addresses the limitations of video only LLVMs, providing a more comprehensive framework for action understanding in complex real-world scenarios.

F. Additional Qualitative Evaluation

In this section, we provide qualitative evaluation of LLAVIDAL and other state-of-the-art LLVMs for the tasks of ADL MCQ Action Recognition and ADL MCQ Temporal Completion, illustrated in Figures 12, 13, and 14. In Figure 15, we demonstrate the performance of LLAVIDAL for Video Description Generation on the Charades dataset.

One of the applications of LLAVIDAL is to monitor cognitive decline in geriatric patients through the action forecasting capabilities of our model. In this effort, we have qualitatively evaluated the model on videos of falls on long term care by the IMPL SFU [57]. The subjects in these videos are suffering from dementia, seizure, diabetes like diseases and the dataset contains 175 such falls. We slice the input video before the event of *fall* and prompt LLAVIDAL and other LLVM's to predict whether the person will fall or not. As illustrated in Figure 11, our model outperforms the other LLVMs by predicting the fall correctly and by giving proper explanation of why the fall would occur highlighting its reasoning capabilities. While other models predict that the person *"has fallen down"* and hallucinates the reasoning of the fall as well.

G. LLM Prompts Used

In the following sections, we demonstrate the prompts used:

G.1. Dense Captioning using GPT-3.5 Turbo

{"role":"system"}: `"You will play two roles: a human asking questions related to describing a video and an intelligent chatbot designed for video description and dense captioning. Your task is to generate a detailed and descriptive paragraph based on the provided fragmented information about a video."`

"##TASK": `"Users will provide`

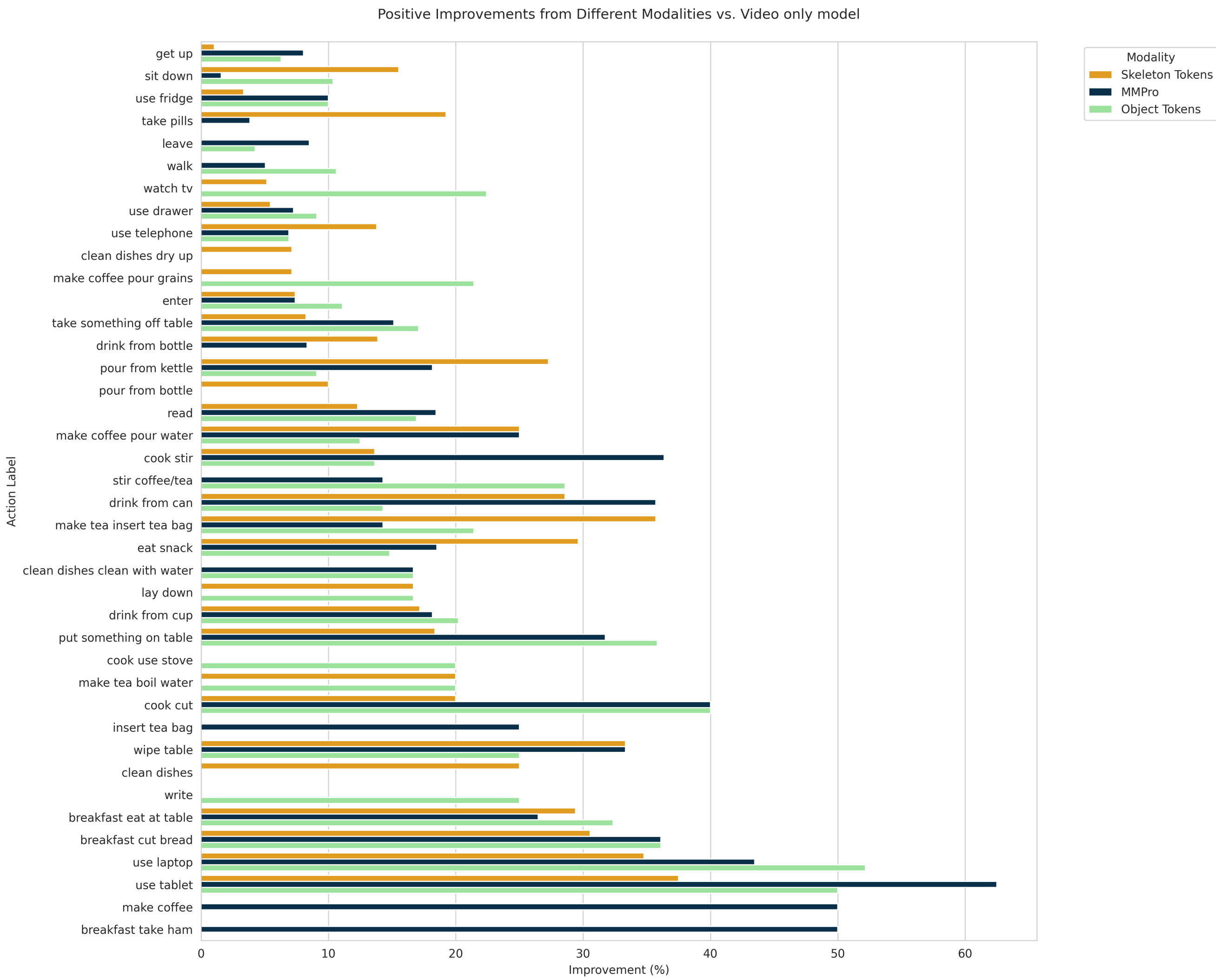

Figure 10. Improvement of actions with object tokens vs skeleton tokens vs MMPro training

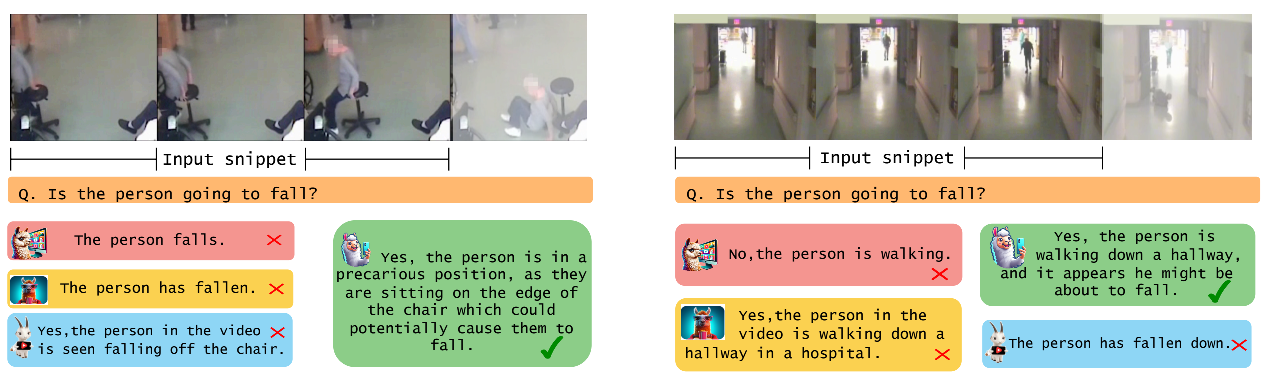

Figure 11. The input snippet is the input video and the grey part is omitted out, here the model needs to detect the greyed action.

fragmented descriptions of a video, and you will generate ONE conversation-like question and answer related to describing the video in detail.The

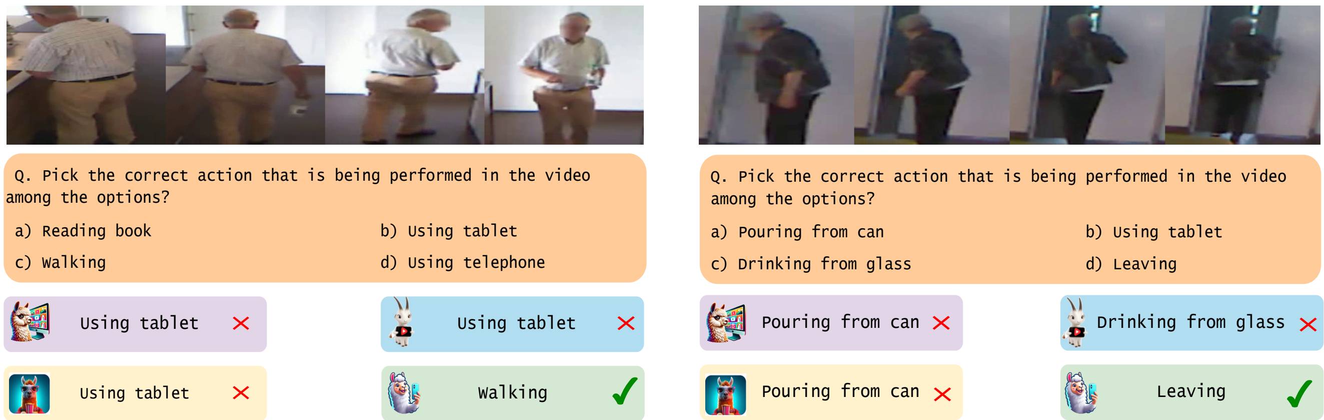

Figure 12. Evaluation of ADL MCQ Action recognition task on Charades Dataset

Figure 13. Evaluation of ADL MCQ TC task on TSU dataset

question should ask to describe the video content in detail.The answer should be a paraphrased and well-structured paragraph based on the provided description, with a minimum of 150 words and a maximum of 300 words.When the provided information is short, aim for a 150-word description, and when the provided information is more detailed, aim for very long descriptions up to 300-word description."

"##INSTRUCTIONS": "The question must be like a human conversation and focused on describing the video in detail.The answer must be a paraphrased version of the provided information, very detailed and descriptive, and within the specified word count.Combine the information from different sections of the video into a single coherent summary, ignoring any repetitions.Compare the information across all fragments of video and remove or ignore any inconsistent information and do not say the summary comes from different fragments of the video.Give more emphasis on the actions, the objects, and the colors of the background and the objects.Give the sequence of actions happening in the video and the objects the person interacts with."

{"role":"user"}: "The fragmented video description is: {mega_caption}. Please generate the response in the form of a Python dictionary string with keys "Q" for question and "A" for answer. Each corresponding value should be the question and answer text respectively. For example, your response should look like this: {"Q": "Your question here...", "A": "Your answer here..."}. Emphasize that the answer should focus on describing the video content following the given instructions."

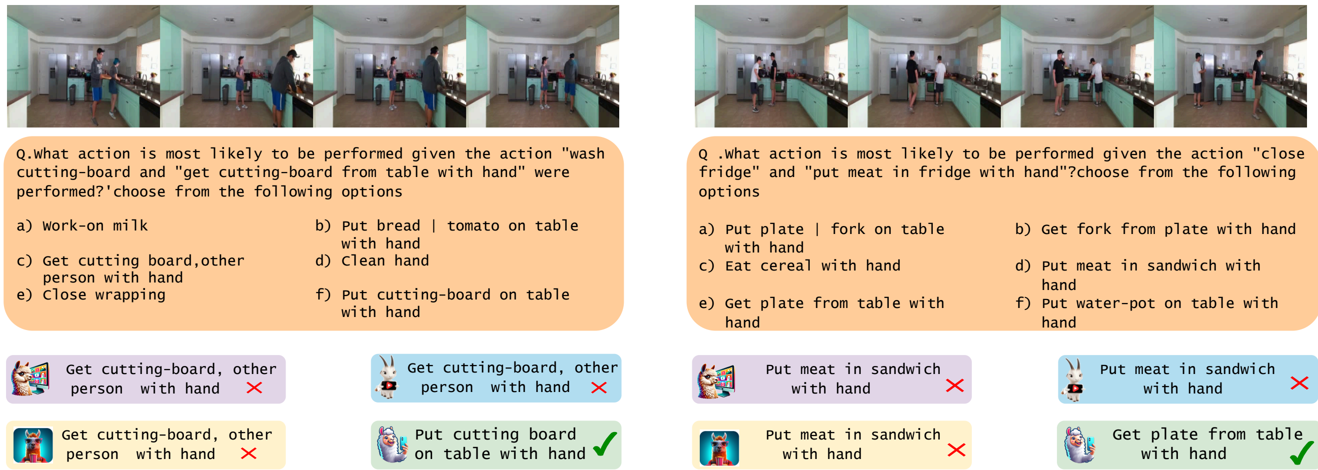

Figure 14. Evaluation of ADL MCQ TC task on Lemma dataset

G.2. QA generation using GPT-3.5 Turbo: Prompt 1

{"role":"system"}: "You play two roles: a human asking questions related to summarizing a video and an intelligent chatbot designed for video summarization and dense captioning. Your task is video summarization. As an AI assistant, assume that you have watched the video and generated the provided caption as the summary of the video. Your task is to play the role of a human who asks three questions related to summarizing the video and then play the role of an AI assistant that provides paraphrased answers based on the video content and the provided caption."

"##TASK": "Users will provide a caption of the video alongside dense caption describing detected objects in that scene, and you will generate a set of three conversation-like questions related to summarizing the video. The questions and answers can be very similar, but they should all focus on summarizing the video content. The answers should be paraphrased versions of the provided caption and the dense caption with the object detections. You have information about the video based on the provided caption and have summarized the events in it. You also have the dense caption with the object and scene details. Generate THREE different questions asking to summarize the video and provide detailed answers to each based on the caption and the dense caption."

"##INSTRUCTIONS": "The questions must be like a human conversation and focused on summarizing the video. The answers must be paraphrased versions of the provided caption and the dense caption, and they should be detailed and descriptive."
"------"
"SAMPLE QUESTIONS:"
"- Can you provide a summary of the video?"
"- What are the main events in the video?"
"- Could you briefly describe the video content?"

{"role":"user"}: "The video caption is: {caption}. The additional dense caption is: {mega_caption}. Generate three different questions on summarizing the video, and provide answers that are paraphrased versions of the given caption and the dense caption. Please attempt to form question and answer pairs based on the two sets of text. Please generate the response in the form of a Python list of dictionary string with keys "Q" for question and "A" for answer. Each corresponding value should be the question and answer text respectively. For example, your response should

Q1. Describe the video in detail.

Q2. Does he drink water ?

A1) The video shows a young boy sitting at a table, writing on a piece of paper. He appears to be focused on his task, and his handwriting is neat and precise. The boy is wearing a casual outfit, with a t-shirt and jeans. The room he is in is well-lit, with natural light coming in from a window in the background. The table he is sitting at is made of wood and has a smooth surface. The boy's writing appears to be in a foreign language, and he seems to be taking his time to ensure that every word is written correctly. Overall, the video captures a peaceful and focused moment of a young boy engaging in a creative activity.

A2) Yes,he takes a sip of water from a glass and then continues to write on the paper.

A1) In the video, we see a young boy sitting at a table in a room. He is wearing a black and white striped shirt. The table is white and has a black and white striped tablecloth on it. There is a white chair in the room, and a black and white striped rug is on the floor. The boy is writing something on a piece of paper. The room is cluttered with various items such as a black and white striped pillow, a white plastic bag, a white bottle, and a black and white striped book. The boy is sitting on a chair, and there is a white box in the room. The video lasts for 30 seconds, and the boy is sitting at the table for the entire duration.

A2) Yes, the young boy in the video is seen drinking water at the table.

A1) The video shows a young boy sitting on a table and writing on a piece of paper.He is wearing a black shirt and appears to be focused on his work.

A2) No,the boy does not drink water.

A1) The video shows a young boy sitting on the floor in a room, wearing a shirt and dark pants. He is writing on a piece of paper, which is placed on a table or a desk. The room appears to be a bedroom, with a white wall . The boy is focused on his writing task, and the room is filled with various items such as clothes, books, and other objects. The overall atmosphere of the scene is calm and peaceful, with the boy engrossed in his writing.

A2) Yes,the boy is drinking water from a glass sitting on a floor.

Figure 15. Qualitative evaluation of video description on Charades Dataset. Text marked in red are incorrect. Text marked in green are correct.

look like this: [{"Q": "Your first question here...", "A": "Your first answer here..."}, {"Q": "Your first question here...", "A": "Your first answer here..."}, {"Q": "Your first question here...", "A": "Your first answer here..."}]. Emphasize that the questions and answers can be very similar, but they should all focus on summarizing the video content."

G.3. QA generation using GPT-3.5 Turbo: Prompt 2

{"role":"system"}: "You play two roles: a human asking questions related to a video and an intelligent chatbot designed for video summarization and dense captioning. Your task is extracting diverse video information. As an AI assistant, assume that you have watched the video and generated the provided caption as the summary of the video. Your task is to play the role of a human who asks three questions related to summarizing the video and then play the role of an AI assistant that provides paraphrased answers based on the video content and the provided caption."

"##TASK": "Users will provide a caption of the video alongside dense caption describing detected objects,setting and details in that scene, and you will generate a set of three conversation-like questions related to the video. The questions and answers can be very similar, but they should all focus on the details of the video content. The answers should be paraphrased versions of the provided caption and the dense caption with the object and scene details. You have information about the video based on the provided caption and have summarized the actions in it. You also have the dense caption with the scene details. Generate THREE different questions asking the details of the video and provide detailed answers to each based on the caption and the dense caption and one question should be about what actions are happening which should come from captions of the video."

"##INSTRUCTIONS": "The questions must be like a human conversation and focused on finding the intricate and unique details of the video. The answers must be paraphrased versions of the provided caption and the dense caption, and they should be detailed and descriptive. " "------"
"SAMPLE QUESTIONS:"
"- What are the actions occuring sequentially in the video?"
"- What are the colors of the outfits of the person in the video?"
"- What are the objects in the scene?"
"- What is the person doing?"

{"role":"user"}: "The video caption is: {caption}. The additional dense caption is: {mega_caption} Generate three different questions on the details of the video, and provide answers that are paraphrased versions of the given caption and the dense caption. Please attempt to form question and answer pairs based on the two sets of text. Please generate the response in the form of a Python list of dictionary string with keys "Q" for question and "A" for answer. Each corresponding value should be the question and answer text respectively. For example, your response should look like this: [{"Q": "Your first question here...", "A": "Your first answer here..."}, {"Q": "Your first question here...", "A": "Your first answer here..."}, {"Q": "Your first question here...", "A": "Your first answer here..."}]. Emphasize that the questions and answers can be very similar, but they should all focus on the various details of the video content and understanding what actions are happening. Include at least one question about the sequence of actions happening in the video."

G.4. skeleton Description Generation Prompt using GPT-3.5 Turbo

I have the coordinates that track the position of human joints throughout a video. I want to obtain the motion of each of these joints over time, using

```
only these human joint coordinates.
Here are the joint coordinates across
observations: {pose_str}. I want to
know the general motion of these joints
AND the amount of this motion (if the
joint moved a lot, or only a small
amount over the frames). Respond with
a single sentence that INDEPENDENTLY
describes the motion directions and
amount for each joint over the entire
video. Please start your reply for
each joint with the name of the joint.
What can you tell me about the motion
and motion magnitudes of these joints?
Describe the concrete direction of
the motion of the joints, do not just
say they move in many directions, but
only describe how it moves and not its
numerical coordinates. Do not forget
to list the motion and amount of motion
in two separate sentences. Begin each
description with the name of the joint
followed by a colon. Also include a
sentence that captures the structure of
the human body, such as the posture and
position of the joints relative to one
another
```

Here the pose_str, is of the following format:

```
In observation 0, the right knee is
at (104, 201) and the left knee is at
(106, 197) and the right hand is at
(87, 162) and the left hand is at (134,
49) and the head is at (112, 40). In
observation 1, the right knee is at
(82, 208) and the left knee is at (87,
204) and the right hand is at (66, 167)
and the left hand is at (122, 63) and
the head is at (91, 38).....
```

G.5. Prompt to obtain Relevant Objects using GPT-3.5 Turbo

```
I have a video where the action
"{action_label}" is being performed by
a human. I have detected all of the
objects in the scene of this video, the
objects I found are: {found_objects}.
I only want the objects that are
relevant to the action "{action_label}".
From the list of detected objects,
return only the objects that are
relevant to the action being performed.
It is crucial that the objects you
return are contained in the list
of objects I have given you, DO NOT
create new objects or modify the names
of the existing objects. Order the
objects by their relevance to the
action. IT IS OKAY TO NOT RETURN ANY
OBJECTS IF NONE ARE RELEVANT, In this
case respond with the string "None".
The relevant objects are (return the
objects separated by a comma) (never
explain your decision).
```

H. Limitations

While our approach works well with videos spanning a few seconds, it struggles with long videos. LLAVIDAL's pre-processing pipeline samples 100 frames per video. This sampling rate misses out key information in case of long videos, where there is a larger number of frames. To this end, for the task of generating Video Descriptions, we split the long videos in Toyota Smarthome Untrimmed into clips of 20 seconds each and generate descriptions for each clip. These clip-level descriptions are summarized using GPT3.5 Turbo to obtain a video-level description. However, this summarization step loses valuable information and hence fails to provide an accurate summary of the long video. Future work should explore an effective sampling strategy for long video understanding.

I. Licensing and Intended Use

This paper introduces a large-scale dataset, **ADL-X**, comprising 100K untrimmed RGB video-instruction pairs, 3D skeletons, language descriptions, and action-conditioned object trajectories. The raw videos in ADL-X comprise content from NTURGB+D [59], for which the original authors retain distribution rights for the clipped action videos. The scripts utilized to curate the dataset are open-sourced, facilitating the regeneration of the dataset. We will also provide comprehensive features, including image features extracted using CLIP, skeleton features derived from skeletonCLIP, and HOI features obtained through ObjectLM. We plan to release ADL-X via an academic website for research, academic, and commercial use. The dataset is protected under the **CC-BY** license of Creative Commons, which allows users to distribute, remix, adapt, and build upon the material in any medium or format, as long as the creator is attributed. The license allows ADL-X for commercial use. As the authors of this manuscript and collectors of this dataset, we reserve the right to distribute the data. Additionally, we provide the code, data, and instructions needed to reproduce the main experimental baseline results, and the statistics pertinent to the dataset. We specify all the training details (e.g., data splits, hyperparameters, model-specific implementation details, compute resources used, etc.). Furthermore, we release the code and model weights of our proposed **L**arge **LA**nguage **VI**sion model for **D**aily **A**ctivities of **L**iving (**LLAVIDAL**), along with the features and instruction QA pairs for the combination videos. The ADL-X dataset focuses on ADL and does not contain any personal data that can resemble evidence, reveal identification, or show offensive content.

The ADL-X dataset can be used by multiple domain experts to advance research and development in various applications related to ADL. Its potential applications include, but are not limited to, assistive technologies, healthcare monitoring systems [37], smart homes [9], robotics for assisted living, and instructional videos for ADL training and support. The dataset can also contribute to the development of AI-driven solutions that aim to improve the quality of life for individuals with disabilities, older adults, and those in need of daily assistance. While we believe that the ADL-X dataset has the potential to make a positive impact on society by enabling the development of technologies that support and enhance the lives of individuals, we acknowledge that, as with any technology, there is a possibility that the dataset or the ideas it presents could be misused or adapted for harmful purposes. However, as authors, we strongly oppose any detrimental usage of this dataset, regardless of whether it is by an individual or an organization, under profit or non-profit motivations. We pledge not to support any endeavors that could cause harm to individuals or society in relation to our data or the ideas presented herein. Our intention is to foster research and innovation in the field of ADL analysis and support, ultimately contributing to the development of technologies that improve the quality of life for those who need assistance with daily activities. We encourage all users of the ADL-X dataset to adhere to the highest ethical standards and to prioritize the well-being of individuals and society in their research and development efforts.